DOCUMENT RESUME

ED 336 428 TM 017 270

AUTHOR	Kane, Michael T.
TITLE	An Argument-Based Approach to Validation.
INSTITUTION	American Coll. Testing Program, Iowa City, Iowa.
REPORT NO	ACT-RR-90-13
PUB DATE	Dec 90
NOTE	49p.
AVAILABLE FROM	ACT Research Report Series, P.O. Box 168, Iowa City, IA 52243.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Evaluation Methods; Formative Evaluation; *Inferences; Models; Predictive Measurement; *Research Methodology; Research Needs; *Scores; Summative Evaluation; Test Interpretation; *Test Validity
IDENTIFIERS	Argument Research; Validation Verification and Testing Techniques

ABSTRACT

The literature on validity provides much more guidance on how to collect various kinds of validity evidence than it does on which kinds of evidence to collect in specific cases. An argument-based approach to validation redresses the balance by linking the kinds of evidence needed to validate a test-score interpretation to the details of the interpretation. This approach can be summarized in terms of a six-step iterative process; the first three steps constitute the formative stage and the last three steps constitute the summative stage. The interpretation is defined as an interpretive argument leading from test scores to statements and/or actions and is validated by evaluating the plausibility of this argument. The evidence supporting the interpretive argument constitutes an argument for the validity of the corresponding interpretation. The details of this validity argument depend on the specific inferences and assumptions in the interpretive argument, but the process of evaluating the interpretive argument provides a general, argument-based approach to validation. A 40-item list of references and a glossary of selected terms are included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED336428

ACT Research Report Series 90-13

An Argument-based Approach to Validation

Michael T. Kane

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. FERGUSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

December 1990

TM017270

ACT

BEST COPY AVAILABLE

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

© 1990 by The American College Testing Program. All rights reserved.

An Argument-based Approach to Validation

Michael T. Kane

ABSTRACT

The literature on validity provides much more guidance on how to collect various kinds of validity evidence than it does on the kinds of evidence to collect in specific cases. An argument-based approach to validation redresses the balance by linking the kinds of evidence needed to validate a test-score interpretation to the details of the interpretation. The interpretation is defined as an interpretive argument leading from test scores to statements and/or actions and is validated by evaluating the plausibility of this argument. The evidence supporting the interpretive argument constitutes an argument for the validity of the corresponding interpretation. The details of this validity argument depend on the specific inferences and assumptions in the interpretive argument, but the process of evaluating the interpretive argument provides a general, argument-based approach to validation.

An Argument-based Approach to Validation

According to the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1985), validity "...refers to the appropriateness, meaningfulness, and usefulness of specific inferences made from test scores". Messick's (1989) definition emphasizes the appropriateness of score-based actions in addition to the appropriateness of inferences:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. (p. 13)

Cronbach (1971, p. 443) defined validity in terms of "the soundness of all the interpretations of a test...". All three of these definitions relate validity to the appropriateness of the inferences included in test-score interpretations. However, the implications, in terms of the data to be collected, the analyses to be performed, and the arguments to be made, of linking validity with interpretations have not been fully developed.

The interpretation of test scores involves inferences from the test scores to various conclusions, and possibly, to decisions about appropriate actions. The conclusions may include statements about persons or groups, predictions of future performance for the person or group, or explanations of observed behavior. The reasoning leading from a score to one or more such conclusions is necessarily based on assumptions. Justifications for possible actions based on test scores involve additional inferences and assumptions, including assumptions about the relative values of the different possible outcomes of various actions.

The inferences and assumptions constitute an argument, leading from the test scores to the statements, predictions, explanations, or decisions included in the interpretation. The argument might be based on scientific models or on pragmatic concerns. It might be presented formally or informally. It might be supported by theory, by empirical research, and/or by appeals to "common sense." It might be stated in detail or only sketched. The nature of the argument might vary along a number of dimensions, including content, level of detail, and mode of presentation. In any case, justification is required for the claim that certain kinds of statements can be made or certain actions are appropriate based on test scores, and justification is provided in the form of argument. Proposed interpretations and uses are valid to the extent that the reasoning involved in the interpretation is sound, reasonable, plausible - that is, valid.

The argument-based approach to validation adopts the interpretation as the framework for collecting and presenting validity evidence and explicitly associates validity with the plausibility of the various assumptions and inferences involved in the interpretation. Treating validation research as an effort to evaluate the inferences and assumptions inherent in test-score interpretations provides a clear framework for evaluating the validity of interpretations assigned to test scores. Furthermore, because it focuses on the details of the argument inherent in the interpretation, this approach also has potential for improving test design and use, rather than simply documenting successes and failures.

This essay begins by identifying a common weakness in discussions of validity--the lack of explicit guidelines for selecting the types of evidence to be employed in validating test-score interpretations. As a potential solution to this problem, test-score interpretations are analyzed in terms of

the arguments associated with the interpretations, and validity is defined in terms of the overall justification for these arguments.

The argument-based approach has a pragmatic emphasis. Validation research is assumed to involve a systematic effort to improve (1) the accuracy of conclusions based on test scores, (2) the appropriateness of the uses made of these scores, and (3) the quality of the data-collection procedures designed to support the proposed conclusions and uses.

A Lack of Guidelines for Validity Evidence

According to the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1985, p. 9), "Validity is the most important consideration in test evaluation." However, like many virtues, validity is more honored than practiced. Ebel's assertion (1961) made over 25 years ago still rings true:

> Validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few. Test validation, in fact, is widely regarded as the least satisfactory aspect of test development. (p. 640)

More recently, Messick (1988, p. 34) has pointed out a "persistent disjunction between validity conception and validation practice", because the "conception" always requires multiple lines of evidence, but, in practice, validation evidence is often very limited. Feldt and Brennan (1989, p. 143) have suggested three reasons why "test theorists and researchers seem to devote an inordinate amount of attention to the reliability of measures as compared with validity": (1) the mathematical rigor of theories of error used to analyze reliability, (2) the fact that reliability depends on test data alone, and (3) the importance of subjective judgment in the study of validity.

Validity has proven to be an elusive concept. Thinking about validity can be frustrating, and trying to do something about validity can be even more frustrating. One reason for a high level of frustration in trying to validate a test-score interpretation is the lack of clear guidelines for what needs to be done to validate a test-score interpretation. By contrast, generalizability theory provides relatively clear guidelines for the kinds of analyses required to support the generalizability of test scores. In particular, if the intended interpretation suggests generalization over certain facets, generalizability theory requires that the sampling error associated with these facets be evaluated to insure that it is not too large (Cronbach et al., 1972; Kane, 1982; Brennan, 1983; Feldt & Brennan, 1989). There is no parallel process for systematically examining validity.

Twenty years ago, Cronbach described the basic problem as it applied to construct validity and outlined a solution based on an explicit statement of the proposed interpretation of the construct:

> The most serious criticism to be made of programs of construct validation is that some of them are haphazard accumulations of data rather than genuine efforts at scientific reasoning. Merely to catalog relations between the test under study and a variety of other variables is to provide a do-it-yourself kit for the reader, who is left to work out his own interpretative theory. Construct validation should start with a reasonably definite statement of the proposed interpretation. That interpretation will suggest what evidence is most worth collecting to demonstrate convergence of indicators. A critical review in the light of competing theories will suggest important counterhypotheses, and these also will suggest data to collect. Investigations to be used for construct validation, then, should be purposeful rather than haphazard. (1971, p. 483)

In spite of this prescription, specific guidance on how to validate test-score interpretation is not very evident in the current literature on validity. Although the concept of validity has been analyzed in some detail, the strategies proposed for validating specific test-score interpretations tend to

be somewhat vague and general. For example, in the introduction to Chapter 1, on validity, the most recent Standards (AERA, et al., 1985) state that:

> An ideal validation includes several types of evidence, which span all three of the traditional categories. Other things being equal, more sources of evidence are better than fewer. However, the quality of the evidence is of primary importance and a single line of solid evidence is preferable to numerous lines of evidence of questionable quality. Professional judgment should guide the decisions regarding the forms of evidence that are most necessary and feasible in light of the intended uses of the test and any likely alternatives to testing.
>
> Resources should be invested in obtaining the combination of evidence that optimally reflects the value of a test for an intended purpose. In some circumstances, evidence pertaining to test content is critical; in others, criterion-related evidence is critical. Evidence regarding the psychological meaning of the construct is usually relevant and may become the central issue. (p. 9)

Although the first paragraph in this passage states that more evidence is better than less, and that the quality of the evidence is important, it does not specify the kinds of evidence to use. The suggestions for selecting particular kinds of evidence (e.g., the use of "professional judgments" and the desirability of "obtaining the combination of evidence that optimally reflects the value of a test for an intended purpose"), while sensible, are very general. The second paragraph does address the issue of relevance, but no specific criteria are provided for deciding when to emphasize a particular kind of evidence.

The first standard in Chapter 1 of the Standards (AERA, APA, NCME, 1985) links the choice of evidence for validity to the "major types of inferences" being recommended.

> Standard 1.1...Evidence of validity should be presented for the major types of inferences for which the use of a test is recommended. A rationale should be provided to support the particular mix of evidence presented for the intended uses. (Primary)

> Comment:
> Whether one or more kinds of validity evidence are appropriate is a function of the particular questions being asked and of the context and extent of previous evidence.

Notice that the nature of the "rationale" to be provided is left unspecified. The other standards in Chapter 1 deal with technical issues involved in designing studies of various kinds, but do not provide additional guidance on the more fundamental questions of what kinds of studies to conduct.

Messick (1988) has criticized the Standards (AERA, APA, NCME, 1985) for accepting the idea, presented in the comment on Standard 1.1, quoted above, that different validation efforts might involve different types of evidence. Messick (1988, p. 35) maintains that this comment

> leaves the door open for an interpretation that there exist circumstances under which only one kind of validity evidence--be it content-related, for example, or criterion-related--may be adequate and fitting for a particular applied purpose. This selective reliance on one kind of validity evidence, when it occurs, is tantamount to reliance on one kind of validity as the whole of validity, regardless of how discredited such overgeneralization may have become...

Messick is concerned that the wording of the Standards might encourage reliance on very limited evidence for validity. He finds reasons in current practice for his concerns about too much flexibility in the Standards:

> A pessimist might view the current state of testing practice as blatant hypocrisy, because of the inconsistency between expressed principles of unified validity on the one hand and widespread behavior of selective reliance on limited kinds of validity evidence on the other (Messick, 1988, p. 36).

Messick has stated his views particularly forcefully, but he is not alone in being concerned about reliance on very limited kinds of validity data (e.g., see Cronbach, 1971, 1989; Guion, 1977, 1980; Tenopyr, 1977; Angoff, 1988).

The lack of specific guidelines for identifying the kinds of data that are most relevant to the validity of a proposed test score interpretation

poses two serious risks. First, the absence of guidelines encourages the practice of selecting evidence to be used for validation mainly on the basis of convenience (in the worst case, picking one "easy" kind of evidence and treating that as a complete answer to the question of validity). The validator is told "more sources of evidence are better than fewer" with little guidance on the sources to be preferred. Given these guidelines and the inevitable limitations on available resources, it would make sense to use the most easily-collected data, since this would facilitate the collection of data from many sources.

Second, the lack of guidelines for deciding on the relevance of different kinds of evidence makes it difficult, if not impossible, to develop clear criteria for how much progress has been made at any given point. An essentially infinite range of studies could be relevant to the validity of an interpretation; if no distinctions are made about the degree of relevance, it is not clear that any limited set of studies could be considered adequate, or even to represent substantial progress. This lack of criteria for gauging progress may reinforce the "persistent dysfunction between validity conception and validity practice," (Messick, 1988, p. 34) by limiting the effectiveness of both the test developer's sense of satisfaction in doing a good job (the carrot) and the effectiveness of external standards (the stick) in encouraging greater effort on validation.

There are, of course, two important sources of guidance for judging the relevance of validity evidence, one implicit, and one explicit but somewhat limited. The implicit source of guidance consists of the specific types of studies and methods of analysis discussed under the heading of validity. By focusing on certain types of data and analyses, the literature does implicitly suggest that such data and analyses are particularly relevant to validation

research. For example, of the 25 standards in the chapter on validity (Chapter 1) in the Standards (AERA, APA, NCME, 1985), 16 address procedures for generating or reporting criterion-related evidence (1.5 and 1.11-1.25), suggesting that criterion-related evidence is potentially an important part of validation research. There are two standards (1.6, 1.7) on content-related evidence, and three (1.8-1.10) on construct-related evidence. The Standards provide more guidance on how to collect and analyze various kinds of validity evidence, than on how to choose among the large number of options available.

The more explicit guideline for selecting types of evidence springs from the important insight that the validity of a test-score interpretation depends on the challenges that can be leveled against the interpretation (Cronbach, 1971, 1988; Messick, 1988, 1989). Adopting this approach, validation studies would seek to evaluate the most serious challenges to the validity of a proposed interpretation. But a reliance on the investigation of plausible, rival hypotheses lacks a positive focus for developing a validation effort (what Lakatos, 1978, calls a positive heuristic) and does not, in itself, provide criteria for assessing the seriousness of various challenges. Since there are potentially an infinity of possible challenges to any interpretation, the validator may be put into the position of simply reacting to the loudest and most persistent challenges. While it is important to accept well-founded criticism and to react to reasonable challenges, this in itself, is not enough. A positive case for the plausibility of the interpretation is needed.

The existence of very specific guidelines for how to conduct certain kinds of studies along with relatively weak guidelines for what kinds of studies to include in a validation effort leads to a situation analogous to that of the airline passengers in an old joke. The pilot has good news and bad news. "The bad news is that a storm has knocked out our radio and compass

and we are completely lost. The good news is that we have a strong tail wind and are making excellent time." To be coherent and relevant, any research program needs a compass, a basis for selecting the questions to be addressed and for setting priorities among these questions. In the case of scientific research, theories play a central role in defining the research agenda; in validation, the inferences and assumptions inherent in the proposed interpretation define the research agenda.

Interpretations as Arguments

The analysis of validation presented in this paper is based on two kinds of arguments. The first of these, the interpretive argument includes the assumptions and inferences involved in the interpretation of the test scores. Interpretive arguments embody the reasoning leading from the test scores to statements about some object of measurement and possibly to decisions.

The interpretive argument contains a number of inferences and assumptions (as all arguments do). The data to be gathered in validation studies are those that are most relevant to the inferences and assumptions in the specific interpretive argument under consideration. It is the content of the interpretive argument that determines the kinds of evidence needed for validation. The interpretive argument also provides a basis for identifying the most serious challenges to a proposed interpretation--challenges that expose weaknesses (e.g., hidden assumptions) in the interpretive argument.

The validity of an interpretation can be defined in terms of the degree to which the interpretive argument is plausible and appropriate. To validate the interpretation is to provide convincing evidence that the interpretive argument is sound, reasonable, plausible (or "valid" in the sense that an

argument is valid). In marshalling evidence to support the interpretive argument, we are, in effect creating a new argument, the validity argument.

The validity argument presents the case for believing the interpretive argument, strong or weak as it may be. The validity argument evaluates the plausibility of the proposed interpretive argument, and can be viewed as a meta-argument, relative to the interpretive argument. The interpretive argument makes the interpretation more explicit, and the validity argument justifies the interpretation. As an aid in keeping this terminology straight, a brief glossary is included in Table 1.

A major advantage of an argument-based approach to validation is that it provides guidelines for choosing the most appropriate kinds of evidence in particular cases. The kinds of evidence that are most important in developing a sound validity argument for a proposed interpretation are those that support the assumptions made in the interpretive argument, particularly those parts of the interpretive argument that are most problematic, a priori. The interpretive argument provides a clear basis for choosing the kinds of evidence to be included in the validity argument.

An Example

To make the development more concrete, it may be useful to sketch an example that is relatively simple, yet illustrates the central points. Suppose we have a sequence of college mathematics courses including the regular first course in the sequence, calculus, and a remedial algebra course for students who are not adequately prepared to take calculus. Assume further that we are going to use an algebra test to "place" students into one of these two courses. Our example represents a particularly simple case of placement testing, (Sawyer, 1989; Frisbie, 1982; Willingham, 1974; Cronbach & Gleser, 1965).

On one level, the interpretation is quite simple. The test scores are interpreted as a measure of competence in algebra and as a measure of "readiness" for the regular course. Students who score at or above some cutoff presumably have learned enough algebra to be considered prepared for the regular course and are assigned to it; students with scores below the cutoff are considered unprepared for the regular course and are assigned to the remedial course.

However, even in this simple case, laying out the interpretive argument can get quite complicated. Because this example is intended simply to make the subsequent discussion of interpretive arguments and validity arguments a bit more concrete, the interpretive argument will be only sketched.

The interpretive argument for the placement test might go something like this:

(1) Some skill in algebra is needed in order to be successful in the calculus course. That is, skill in algebra is a prerequisite in the sense that students who lack such skill are likely to have great difficulty in dealing with the content of the calculus course.

(2) The placement test measures the algebraic skills required in the calculus course, is reasonably reliable, and is not influenced substantially by any sources of systematic error.

(3) The cutoff score is appropriate in the sense that students with scores at or above the cutoff score have sufficient skill in algebra to succeed in the calculus course, and students who score below the cutoff lack some or all of the algebraic skills needed for the calculus course.

Assumptions (1) and (2) imply that performance on the placement test is relevant to readiness for the calculus course, at least in the sense that

students with low scores on the test are likely to have difficulty in the calculus course, because they lack one of the prerequisites. By adding assumption (3), we can draw the stronger conclusion that students who pass the test (score at or above the cutoff) are "ready" for the calculus course and that students who fail the test are not ready for the calculus course and therefore should be placed in the remedial course.

In sketching the interpretive argument, we have, of course, left out much of the substance of the argument, including a number of important assumptions. For this placement system to be effective, for example, would require that the remedial course be effective in developing the prerequisite knowledge and skill, and that students who have passed the remedial course have an improved chance of succeeding in the calculus course. We would also generally assume that the students who passed the test and were assigned to the calculus course would not benefit substantially from taking the remedial course. These two assumptions represent a special case of what Cronbach and Snow (1977) have called an aptitude treatment interaction. They are often implicit, but they are essential to the logic of the argument.

The appropriateness of the placement system also rests on more fundamental assumptions. For example, we are tacitly assuming that the use of a placement system is preferable to a redesign of the regular course so that the pace and/or sequence of instruction is flexible enough to accommodate all students. We are also assuming that minimizing the number of students who fail the regular course is a sufficiently important goal that it merits the commitment of substantial resources (i.e., a second course and the time and money required for placement testing).

The interpretive argument for the placement test will be developed a bit more in the subsequent discussion, but there will be no attempt to make it

fully explicit. Sawyer (1989) provides a more thorough discussion of the assumptions and inferences involved in placement systems.

Interpretive Arguments

As noted earlier, it is the test-score interpretations that are validated. At their core, interpretations involve meaning or explanation. The first definition of the verb "interpret," in Webster's Ninth New Collegiate Dictionary is: "to explain or tell the meaning of: present in understandable terms." This definition captures much of the meaning of "interpretation" as it is used in discussions of the validity of test-score interpretations. To interpret a test score is to explain the meaning of the score and, thereby, to make at least some of the implications of the score clear.

However, to define "interpretations" in terms of "explanations" and "meanings" is not in itself very helpful in thinking about validation. The specification of meanings and the development of explanations tend to involve considerable difficulty whenever we try to go beyond verbal definitions, which stipulate the meaning of one word or symbol in terms of other words or symbols. It is probably more helpful to examine the structure of interpretations and some of their salient features.

A test-score interpretation always involves an argument, a chain or network of inferences, with the test score as a premise (or premises, in the case of a profile of several scores) and the statements, predictions, decisions, etc. involved in the interpretation as the conclusions. This argument is being referred to as the interpretive argument.

The interpretive argument includes the inferences used in going from test scores to the statements involved in the interpretation and also includes the

assumptions on which these inferences are based. Where test scores are used to make decisions, the reasoning (including assumptions about values) leading to the decision is also part of the interpretive argument.

In interpreting test scores, the conclusions, including proposed actions, are typically stated explicitly. Some intermediate steps in the interpretive argument may also be included explicitly in the interpretation assigned to the test scores. For example, in reporting results for our algebra placement test, a student might be advised that he is not ready to take the calculus course (final conclusion) because he has not mastered some essential skills in algebra (intermediate conclusion). However, most of the interpretive argument is generally left implicit in reporting results and may not be stated explicitly even during test development. Nevertheless, the interpretation entails all of the intermediate assumptions and conclusions involved in going from the test scores to the specific conclusions included in the statement of the interpretation.

The interpretive argument embodies the reasoning that is used (implicitly or explicitly) whenever the interpretation is applied to test scores. The measurement procedure basically assigns a number to some object of measurement. In going from this number to a verbal description of the object of measurement or to a statement (verbal or numerical) about some present or future characteristic of the object, or to a decision of some kind, we are going beyond the scores. The reasoning involved in the interpretive argument may be sound, or it may be faulty. Judgments about the validity of a test-score interpretation are basically judgments about the soundness or plausibility of the interpretive argument (i.e., the validity of the argument).

Inferences and Evidence. The interpretations assigned to test scores generally depend on networks of different kinds of inferences, including generalizations, extrapolations, predictions, causal and noncausal explanations, theory-based inferences, and score-based decisions.

Most, if not all, test-score interpretations involve generalization from the specific observations being made to a broader universe of similar observations. In interpreting scores on the placement test discussed earlier, we generally do not limit our statements to a specific time, a specific place or a specific scorer. In reporting results with sentences like, "John got a 60 on the placement test", rather than the more cumbersome statement, "John got a 60 on the placement test that he took on May 6, in auditorium B, and that was scored by Prof. Jones," we are implicitly assuming that the particular time and place of testing and the choice of scorer are not relevant to the interpretation. We treat the observations as if they have been sampled from some universe of observations, involving different occasions, locations, and observers that could have served equally well; that is, we generalize over some conditions of observations (Cronbach, Gleser, Nanda, and Rajaratnam, 1972; Brennan, 1983). Generalization rests on assumptions about the generalizability of the test scores over some conditions of observation.

Most interpretive arguments also involve extrapolation; conclusions are drawn about behavior that is different in potentially important ways from that observed in the testing procedure. We are likely to interpret scores on our placement test as indicating the ability to use the algebraic techniques covered in the test in a variety of contexts, even though the placement test may consist of discrete, multiple-choice items administered in one testing session. The use of test scores as an indication of non-test behavior assumes that the relationship between the scores and the target behavior is understood

fairly well (Cronbach, 1982; Kane, 1982; Tryon, 1957). The extrapolation may be based on fairly loose notions of similarity or on a detailed analysis of the specific processes used in the two situations (Snow and Lohman, 1984, 1989). Since it is hardly ever the case that we actually draw a random sample from the intended universe, simple generalization is usually not an entirely appropriate model, and there is no sharp distinction between generalization and extrapolation.

Essentially all interpretations also involve, at least implicitly, some theory-based inferences involving possible explanations and/or connections to other constructs. Some interpretations are primarily theory-based, in that the observations in the measurement procedure are of interest primarily as indicators of unobservable constructs. However, even when the focus of the interpretation is more practical than theoretical, theoretical considerations have a role in the interpretive argument. The assumption that the skills measured in the algebra test are prerequisites for the regular calculus course is based on assumptions about processes; we assumed that students would use the concepts and techniques of algebra in solving the problems encountered in the calculus course. The explanations that we incorporate in our interpretations, whether theory-based or common-sense-based, assume the relevance and soundness of the models being employed.

Most educational tests are also linked to some decision. If the test scores were not relevant to any decision, it is not clear why the test would be given. The legitimacy of test use rests on assumptions about the possible outcomes (intended and unintended) of the decision to be made and the values to be associated with these different outcomes (Messick, 1988, 1989).

Each of the inferences in an interpretive argument rests on assumptions that provide justification for the inference. Simple generalizations from the

test score to some domain of behaviors rest on assumptions about the generalizability of observed scores (Cronbach, et al., 1972; Brennan, 1983). Extrapolations are based on assumptions about the relationship (e.g., similarity or overlap in processes used) between the types of behavior actually observed and the types of behavior to which the results are being extrapolated (Cronbach, 1982; Kane, 1982; Snow & Lohman, 1984). Similarly, predictions assume some specific relationship between the test scores and performance being predicted.

Explanations may be based on covering laws, on theories, or on general assumptions about relationships. Any theory-based inferences assume the validity of the theory being used (Cronbach and Meehl, 1955; Cronbach, 1971; Meehl and Golden, 1982). In addition, decisions based on test scores make assumptions about the desirability of various kinds of outcomes, that is, about values (Messick, 1975, 1980, 1981, 1988, 1989; Guion, 1974).

From the point of view of validation, the assumptions are generally the key elements in the interpretive argument. Flaws in the interpretive argument are likely to involve faulty or doubtful assumptions, rather than flaws in logic; errors in logic are generally easier to detect and easier to fix than weaknesses in the assumptions, especially if the argument is not stated clearly. Interpretive arguments cannot be specified with the precision found in logical/mathematical derivations, and are often stated only in the most general terms. Furthermore, the arguments tend to be complex and to involve many assumptions, and, therefore, are difficult to define clearly and to evaluate effectively.

Particularly troublesome are assumptions that are implicit, or "hidden," in the sense of not being explicitly recognized as part of the argument. Hidden assumptions are, of course, a major concern in evaluating any argument,

but they are most likely to cause problems in interpretive arguments that are not stated clearly.

The Validity Argument

The validity argument provides the rationale for accepting the interpretive argument and, therefore, for accepting the interpretation. The validity argument may use new empirical data, the results of previous research, and various kinds of reasoning (ranging from mathematical analyses based on statistical/psychometric models to appeals to common sense) to support various parts of the interpretive argument.

If the validity argument is to support the interpretive argument effectively, it must reflect the structure of the interpretive argument. The interpretive argument represents the reasoning (including inferences and supporting assumptions) inherent in the interpretation and depends on the procedures used to generate test scores and the interpretation being proposed as well as the context in which the scores will be interpreted and used. Therefore, validity arguments are unique in their details but all share a common purpose--to provide a systematic evaluation of the corresponding interpretive argument. To be most effective in checking the interpretive argument, the validity argument should focus on those parts of the interpretive argument that are most doubtful or problematic.

As noted earlier, there are many different types of inferences (e.g., extrapolation, theory-based inferences) and supporting assumptions that may play a role in interpretive arguments. As a result, there are many different types of evidence that may play a role in the validity argument. Since each interpretation tends to involve a network of different types of inferences and

assumptions, a thorough examination of validity would generally involve several types of evidence.

In discussing some of these types of evidence, it is helpful to distinguish two stages in the evaluation of the interpretive argument. The formative stage of the validity argument involves the clarification/explicit definition of the interpretive argument and the development of a preliminary case for the plausibility of the interpretive argument.

The second stage in the evaluation of the interpretive argument involves empirical checks on the assumptions and inferences in the interpretive argument. In this summative stage of the validity argument, a reasonably mature version of the interpretive argument can be subjected to serious, empirical challenges. To the extent that the interpretive argument survives such challenges, our confidence in its validity increases.

The purpose of the formative stage of the validity argument is to layout a preliminary case for the interpretive argument, and the purpose of the summative stage is to subject the interpretive argument to empirical challenges. The distinction drawn here between the two stages is intended to facilitate discussion of the conceptual components of validation research and is not intended to suggest a sharp temporal division.

The use of this terminology parallels the use of the terms "formative" and "summative" in program evaluation. The aim of the formative stage of the validity argument is to develop and refine the interpretive argument, just as the aim of the formative stage of program evaluation is to improve the program. The goal of the summative stage of the validity argument is to arrive at summary judgments about the plausibility of the interpretive argument and, therefore, about the appropriateness of conclusions and decisions being based on test scores, just as the goal of the summative stage

of program evaluation is to determine the effectiveness of the program. The similarity between validation research and program evaluation is no coincidence. Cronbach has explicitly linked the logic of validation to the logic of program evaluation:

> Validation of a test or test use is evaluation (Guion, 1980; Messick, 1980), so I propose here to extend to all testing the lessons from program evaluation. What House (1977) has called "the logic of evaluation argument" applies, and I invite you to think of "validity argument" rather than "validation research" (Cronbach, 1988, p. 4).

House (1980) has pointed out that arguments play a central role in evaluation. In doing so, House (1980, p. 73) emphasized the complexity and lack of certainty in such arguments, suggesting that evaluation "persuades rather than convinces, argues rather than demonstrates, is credible rather than certain, is variably accepted rather than compelling."

The description of validity as "argument" emphasizes the need for various kinds of evidence arranged so that the "argument" as a whole is coherent and convincing. It draws attention to the importance of plausible rival hypotheses. And, it indicates the openness of the enterprise; real arguments about important issues are hardly ever resolved by a simple "yes" or "no" answer. Arguments are plausible or credible, rather than certain.

The distinction between the formative and summative stages in the evaluations of the interpretive argument also parallels Popper's (1965, 1968) distinction between two stages in the development of scientific theories--conjecture and refutation. The interpretive argument can be viewed as a theory or conjecture about the appropriate interpretation for the test scores. In some cases, the interpretive argument may, in fact, be based on a theory implicitly defining a specific construct interpretation, with the theory and the interpretation being tested by the same data (Cronbach and Meehl, 1955), but most interpretive arguments are too loose and "ad hoc" to be

referred to as "theories". Nevertheless, the proposed interpretation, like a theory, can be viewed as a "conjecture" to be developed and defended rather than a fact or a stipulation.

The goal of the formative stage of the validity argument is to develop a plausible conjecture; the goal of the summative stage is to evaluate this conjecture by subjecting it to possible refutation by empirical evidence. If the interpretive argument survives critical analysis and empirical testing, we have a reasonable basis for accepting the interpretation.

The Formative Stage: Developing the Interpretive Argument.

The formative stage of the validity argument involves the development of a plausible interpretive argument for test scores. The developers of a test have some intended interpretation or some intended use in mind even before they begin developing the testing procedure. The initial interpretation may be quite general and/or vague (e.g., we want a placement test to be used in assigning entering college students to mathematics courses) but some goal is needed to get started. The process of developing the interpretive argument is mainly analytic rather than empirical, and involves the specification of the interpretive argument in enough detail so that the assumptions inherent in this argument are clear. Initial judgments about the plausibility of the interpretive argument would be based on the relationship between data collection procedures and the proposed interpretation.

To the extent that the test and the interpretation have been fashioned to be compatible, the interpretation assigned to the test scores tends to be plausible. For example, scores based on the percentage of decisions made by a manager without consulting subordinates could reasonably be interpreted in terms of authoritarianism; an interpretation in terms of authoritarianism

would be much less plausible if the scores were based on how well the manager likes the color green. If the connection between the data collection procedures and the interpretation is not evident, it may be necessary to rely on existing empirical results or theories. For example, the existence of a well supported theory linking authoritarianism with a preference for the color green could make a self-report of color preferences a plausible index of authoritarianism. (Astronomers use the color of a star as an indicator of the temperature, composition, and velocity of the star.)

In the best case, the development and refinement of the interpretive argument would be interwoven with the test development process. The intent would be to construct the test in a way that is consistent with the intended interpretation and to develop an interpretation that makes sense, given the nature of the observations being made and current understanding of the attribute being measured.

The efforts made to build the interpretation into the testing procedures help to make an initial, positive case for the plausibility of the argument linking the test scores and the interpretation. A careful analysis (and documentation) of the test specifications and of the procedures used to develop the test can provide evidence relevant to several aspects of a proposed interpretation. In addition to defining the general domain of content covered by the test, it may be possible to develop some understanding of the types of cognitive processes involved in responding to test items. For example, Nedelsky (1965, Chapter 11) has discussed the characteristics required to support various kinds of interpretations of science items; in particular, Nedelsky (1965, p. 152) suggests that items must present novel situations/problems if they are to measure comprehension rather than simple recall. If process interpretations are to be at all plausible, the test must

be designed in a way that is consistent with the interpretive argument leading from the test scores to conclusions about process. Of course, as Cronbach (1971, p. 453) has pointed out:

> An item qua item cannot be matched with a single behavioral process. Finding the answer calls for dozens of processes, from hearing the directions to complex integration of ideas. The shorthand description in terms of a single process is justified when one is certain that every person can and will carry out all the required processes save one.

That is, analyses of test content and testing procedures cannot, by themselves, establish the legitimacy of process interpretations. Collateral assumptions are clearly necessary if we are to draw inferences about process, and these assumptions have to be justified if the interpretive argument in which they occur is to be accepted. However, content analyses can make an interpretation either more plausible or less plausible. In particular, analyses of test content and testing procedures can sometimes effectively rule out certain interpretations.

A careful analysis of data collection procedures can also reveal possible sources of extraneous variance that may undermine a proposed interpretation. If the test is to be interpreted as a measure of achievement in some domain, then efforts to describe the domain carefully and to develop items that reflect the domain (in terms of content, cognitive level, and freedom from potential sources of systematic errors) tend to support the intended interpretation. An interpretation in terms of a theoretical construct would be facilitated by the use of observations that are related to the construct conceptually and/or that have been linked to the construct empirically.

As part of the test development process, testing materials and procedures may be pilot tested in order to improve the materials and procedures. To the extent that results of pilot testing support the assumption that the test and testing procedures are free of various kinds of possible flaws, these results

support the interpretive argument.

One kind of evidence that would be helpful in validating the mathematics placement system introduced earlier would result from a detailed analysis of the algebraic concepts and techniques actually used in the regular calculus course. Such data could be used to make the interpretive argument more explicit. Instead of assuming that "algebra" is a prerequisite for the regular course, we would claim that specific skills X, Y, and Z are prerequisites. The analyses of the algebraic skills used in the calculus course help us to formulate the assumption about prerequisites more precisely and at the same time provide evidence supporting this assumption. If, in addition, pilot testing data suggests that the test is generally free of flaws and has adequate generalizability, we have a reasonable basis for entertaining the hypothesis that the test measures algebraic skills that are prerequisites for successful performance in the regular course.

In general, then, the formative stage involves the clarification of the interpretive argument and the development of a preliminary positive case for the reasonableness of the interpretive argument, based mainly on existing evidence and the relationship between the procedures used to generate test scores and the intended interpretation. Again, using Popper's (1965, 1968) terminology, we develop a "conjecture" about appropriate interpretations and/or uses of the test scores. Our willingness to take the conjecture seriously is based on the overall plausibility of the interpretive argument and the available evidence for this argument.

The Summative Stage: Empirical Testing of the Interpretive Argument.

The summative stage of the validity argument emphasizes empirical checks on the assumptions in the interpretive argument. The aim is to subject the

interpretive argument to searching criticism by challenging its weakest, most doubtful assumptions. During the formative stage of the validity argument, a preliminary case was made for the plausibility of the interpretive argument. During the summative stage, the validity argument is further developed by subjecting the interpretive argument to empirical challenge.

In order for these challenges to have the greatest benefit, they should involve those parts of the interpretive argument that are most vulnerable. Evidence that provides further support for a highly plausible assumption does not add much to the overall plausibility of the argument. It is the problematic assumptions in the interpretive argument, those that are most subject to doubt, that deserve the most attention. Assumptions can be problematic because of existing evidence indicating that they may not be true, because of plausible alternative interpretations that deny the assumption, because of specific objections raised by critics, or simply, because of a lack of evidence supporting the assumption. The interpretive argument is no stronger than its weakest links, and therefore, the best way to evaluate the argument is to examine its most problematic assumptions.

For the mathematics placement test in our earlier example, the assumption that some level of skill in algebra is a prerequisite for successful performance in a calculus course could be considered unproblematic, especially if the content of the placement test has been explicitly linked to the specific algebraic skills actually used in the calculus course. It would be hard to formulate most problems in calculus without using algebraic notation. Therefore the collection of empirical evidence supporting this assumption would probably not add much to a validity argument.

However, the choice of the cutoff score used in assigning students to different courses is likely to profit from careful scrutiny, depending as it

does on such issues as the relationship between test scores and performance in various courses, the relative losses associated with different kinds of errors, and the implicit selection ratios defined by course enrollment limitations. Evidence indicating that students with scores above the cutoff score generally succeed in the regular course, while students below the cutoff score tend not to succeed, could make a substantial contribution to the validity argument by supporting the choice of cutoff score.

The checking of assumptions in the interpretive argument is likely to bring additional assumptions into play. For example, new assumptions (i.e., substantive and statistical) are made in interpreting the results of any empirical studies. Checking these assumptions will introduce additional assumptions. An effort to check on all assumptions leads to infinite regress, with the number of assumptions to be checked increasing indefinitely. Needless to say, this is not a particularly desirable state of affairs. The solution to this problem that is usually employed in science (Lakatos, 1978; Lakatos & Musgrave, 1970) is to simply take some assumptions as given, as unproblematic background knowledge.

The psychologist who uses electronic timers to measure and record reaction times assumes that the equipment, in working order and properly used, can provide accurate measurements of time. This assumption rests on what we know about physics (the equipment's circuit design, etc.) chemistry (the performance characteristics of alloys, plastics, etc. in the equipment) and astronomy (for the origins of our concepts of time). The interpretation also assumes that the perceptual processes of persons recording and/or interpreting the data are "normal" (e.g., not subject to hallucinations). However, unless there is some specific reason to doubt them, these assumptions are all treated as unproblematic background knowledge. The psychologist may have doubts about

a particular piece of equipment or a particular observer, but is not likely to challenge the basic principles that underlie the interpretation of the observations as representing time intervals.

If the results of the various empirical checks tend to support the assumptions in the interpretive argument, the validity argument is strengthened. The plausibility of the interpretive argument is strongly supported if all of the most problematic assumptions survive searching criticism and empirical evaluation. However, the interpretive argument is always subject to new challenges, and, therefore, the validity of the interpretation is never proven.

If the results of some empirical check indicate that the interpretive argument is flawed, there are several options. Evidence indicating serious problems in the interpretive argument may suggest abandoning the whole enterprise. Alternatively, it may be necessary to make major changes in the intended interpretation or the testing procedures, and therefore to develop a new interpretive argument; major problems might suggest a return to the kinds of analyses employed in the formative stage.

In some cases, it may be possible to solve the problems by making relatively minor modifications in either the interpretation or the testing procedures or both. Such changes may permit the elimination of the assumptions that have been contradicted, while preserving an interpretive argument that serves its basic purpose reasonably well.

If a questionable assumption is not central to the interpretive argument, it may be convenient simply to drop the assumption and, perhaps, thereby limit the interpretation somewhat. For example, suppose that the algebra placement test discussed earlier were also used to place students in science courses. If assumptions about the relation between scores on the test and performance

in science courses turned out to be false, this use of the test could be dropped without weakening the argument for the use of the test scores for placement in mathematics or the more basic interpretation in terms of skill in solving algebra problems. Like scientific theories, test-score interpretations do not necessarily fail because of a single problem; confidence in an established interpretation is more likely to be eroded gradually by a succession of problems than to be overturned by a single "crucial" experiment.

Characteristics of Interpretive Arguments

There are at least five characteristics of interpretations and their associated interpretive arguments that are especially relevant to validity issues. (1) Interpretive arguments are artifacts in the sense that they are created and assigned to the test scores by human beings. They can be developed, revised, or abandoned. They are made, not found. (2) Interpretive arguments are structured, with some assumptions playing relatively basic roles in all of the conclusions and actions based on test scores and other assumptions playing less basic roles. (3) Interpretive arguments are dynamic; they may expand or contract or simply shift their focus. (4) Interpretive arguments may need to be modified to accommodate special circumstances in specific situations. (5) Interpretive arguments are open in the sense that at any given time, they are incomplete and anticipate further development.

(1) The interpretation is an artifact. The interpretation that is assigned to the test scores is not uniquely determined by the observations being made. The possible interpretations for any set of test scores vary along several dimensions, including their focus and their level of abstraction; for example, a test involving passages followed by questions

about the passage could be interpreted, simply, as a measure of skill at answering passage-related questions, as a measure of reading comprehension defined more broadly, as one indicator of verbal aptitude, or as an indicator of some more general construct, such as intelligence. These different interpretations necessarily involve different interpretive arguments.

Because the procedures used to collect data do not uniquely determine the interpretation to be given to results obtained using the procedures, one or more interpretations must be assigned to the test scores. We decide how we will interpret the results of the reading comprehension test. The mathematics placement test discussed earlier was interpreted as a measure of readiness for the regular calculus course, because we chose to use it that way.

Defining the proposed interpretation and specifying the associated interpretive argument are of fundamental importance in evaluating the validity of the interpretation. We validate the interpretation by evaluating the plausibility of the interpretive argument inherent in the interpretation. Some possible interpretations may be highly valid, while others are clearly not valid. In the example given above, the interpretation of the scores in the reading comprehension test in terms of skill at answering passage-related questions is likely to be more solid (although perhaps less interesting) than interpretations involving more general constructs. We cannot evaluate the plausibility of the inferences and assumptions in the interpretive argument very well if we have not identified what these inferences and assumptions are.

Therefore, an important first step in any effort to validate the interpretive argument is to state this argument explicitly. The argument may be changed later, perhaps as a result of validation research, but if the effort to check on the assumptions and inferences in the interpretive argument is to make much progress, the effort needs to begin by stating these

assumptions and inferences fairly clearly. An analogous point is made within the context of generalizability theory where the importance of explicitly defining the universe of generalization proposed for test scores is emphasized (Cronbach, et al., 1972; Brennan, 1983; Kane, 1982).

(2) Interpretive arguments are structured. In the placement-test example, conclusions about skill in solving algebra problems drawn from a certain domain are basic to all of the other interpretations proposed. The conclusions drawn about readiness for the calculus course depend on conclusions about skill in algebra and on additional assumptions (e.g., about the relationship between skill in algebra and performance in the calculus course). Therefore, evidence indicating that the test did not do a good job of measuring skill in algebra (e.g., evidence that the test made inordinate demands on reading skills) would tend to cast serious doubt on conclusions about readiness for the calculus course. However, empirical evidence indicating that the test was not a very good indicator of readiness for the calculus course (perhaps because the calculus instructor teaches the algebra needed for the calculus course) would not necessarily cast doubt on the interpretation of test scores in terms of skill in algebra. There is a definite lack of symmetry here. Within the interpretive argument proposed for the placement test, score-based conclusions about skill in algebra are basic to the other parts of the argument. By contrast, assumptions about the utility of the test for placement in any sequence of courses apply only to certain uses of the test, and are therefore less basic.

Because of their structure, interpretive arguments do not have to be accepted or rejected as a whole; we can change or reject parts of the argument while retaining other parts. In particular, specific assumptions and inferences that do not support other inferences and assumptions might be

altered without much change in the general shape of the interpretive argument. Therefore questions about the validity of an interpretation, that is questions about the plausibility of an interpretive argument, do not generally lead to a yes-or-no answer.

(3) Interpretive arguments are dynamic. As new information becomes available, the interpretive argument may expand to include new types of inferences. Empirical results may support generalization to a wider domain or extrapolation to a new domain. Conversely, new results may tend to refute assumptions that supported part of an interpretive argument, thus forcing a narrower interpretation. Society's priorities and/or values may change, leading to changes in how test scores are used.

As research proceeds, deeper or more sophisticated explanations for the test scores may be developed; for example, a process model describing how students solve algebra problems could greatly expand the scope and depth of the interpretation given to our placement test. Similarly, the development of new theoretical approaches to reading is bound to influence our interpretation of scores on a reading comprehension test. Similarly, Nagel (1971), Meehl (1950), and Lakatos (1978) have described the dynamic nature of scientific theories and of the concepts embedded in these theories.

The malleability of interpretations can make validation more difficult or easier. A changing interpretation presents the validator with a moving target. However, it is also possible, in many cases, to make some adjustments in the intended interpretation, based on validity data. That is, we can sometimes strengthen the case for the validity of the interpretive argument by changing the interpretive argument to fit the data. It will be argued later that one possible criterion for evaluating validation research would be the extent to which the research improves the interpretation by making it clearer,

more solidly based, and more accurate, and improves the test by eliminating flaws and sources of error.

(4) The general form of the interpretive argument may need to be adjusted to reflect the needs of specific examinees or to reflect specific circumstances that might have an impact on the test scores. The general version of the interpretive argument, which is used in developing the validity argument, is intended to apply to some population of examinees and cannot take explicit account of all of the special circumstances that might affect an examinee's performance. In applying the general version of the argument to an examinee, we assume that the examinee is drawn from appropriate population and that there are no circumstances that might alter the interpretation. To the extent that this assumption is not plausible in a specific case, we may need to adjust the interpretive argument, the validity argument, or both for that case.

Such adjustments may be made for subpopulations and for individuals. For example, within the subpopulation of examinees with a specific handicap, the interpretive argument may need to be adjusted to reflect the impact of the handicap (see Willingham, 1988); the interpretive argument will change and therefore the validity argument will change. If testing procedures are adjusted to accommodate the needs of a handicapped student, it may be necessary to add evidence supporting the comparability of scores obtained under special testing procedures to the validity argument (Willingham, 1988, p. 98). The general form of the interpretive argument may also need to be modified for individual examinees to reflect special circumstances (e.g., due to illness, lack of motivation).

Interpretive arguments make many assumptions that are unproblematic under ordinary circumstances (e.g., that examinees can hear instructions that are

read to them), but that may be problematic for specific examinees (e.g., hearing impaired examinees) or under special circumstances (a noisy environment). The assignment of an interpretation to a specific test score is an instantiation of the general form of the interpretive argument. The reasonableness of the resulting specific interpretive argument depends on the reasonableness of the general form of the interpretive argument and on the extent to which the interpretive argument applies to the specific situation under consideration.

(5) Interpretive arguments are open. They tend to be somewhat fuzzy around the edges. Initially, the intended interpretation is likely to be stated in very general terms, for example, in terms of "reading comprehension" or "readiness" for a particular course. The interpretive argument is then correspondingly loose. During the formative stage of the validity argument, the interpretive argument is developed and made more explicit. However, even the most highly developed interpretive arguments do not attain the precision of mathematical derivations; rather they are combinations of some theory, some logic, and general arguments for the plausibility of assumptions and inferences.

Therefore, the evaluation of the interpretive argument (i.e., the validity argument) does not typically involve a simple, valid/invalid decision, as it might in logic or mathematics. The validity argument is necessarily judgmental, leading to conclusions about the degree of validity, or plausibility, of the interpretive argument rather than a simple yes/no decision.

In general, then, interpretations and interpretive arguments are artifacts developed by human beings, they have structure, they change with time, they may need to be modified for particular examinees or circumstances,

and they are open in the sense that they could always benefit from additional work. The details of the argument depend on the test development and test administration processes, the types of statements/conclusions and decisions proposed for the test scores, and the context or situation in which the data are generated and used. As a result, each interpretive argument is unique and the evidence needed to support the interpretive argument, that is the validity argument, is also unique.

A Six-Step Process

The argument-based approach to validation can be summarized in terms of a six-step iterative process with the first three steps constituting the formative stage and the last three constituting the summative stage. The approach is open-ended in the sense that there would always be more work that could be done, but it does provide a definite place to begin and criteria for choosing what to do next at each stage in the inquiry.

Note that the six-step process presented here assumes that serious work on validation begins with the development of the testing procedure. In practice of course, it is often necessary to evaluate the validity of specific interpretations assigned to existing testing procedures and therefore the opportunities to adjust the testing procedure may, in practice, be limited.

This six-step process is not intended as a checklist or a cookbook to be used in conducting validation studies. Each validity argument needs to be tailored to the corresponding interpretive argument. Rather, the six-step process is intended to outline the argument-based approach as clearly as possible without getting into specific examples.

Step 1: Specify the interpretation by stating the interpretive argument as clearly as possible. The first step requires the development of the

interpretive argument, or, if a vague argument already exists, its clarification. Of particular interest at this point is the identification of the specific inferences being made, and the identification of the assumptions needed to support these inferences.

Step 2: Evaluate the plausibility of the interpretive argument by examining the reasonableness of its assumptions and inferences. In addition to evaluating the general coherence of the argument, the plausibility of each inference and supporting assumptions would be examined. In some cases, it may be possible to check the assumptions against relevant, previous research and any new data collected while developing the measurement procedures (e.g., item-analysis data and the results of generalizability studies). In other cases, the evaluation of assumptions at this step in the process would be based mainly on judgment and general experience.

Step 3: Make any changes suggested by the evidence. At one extreme, the evidence may simply support the interpretive argument as formulated in Step one. At the other extreme, the evidence may be so damaging that the interpretation is basically untenable and the whole enterprise is abandoned. More generally, the evidence will suggest some changes in the interpretive argument or the testing procedures. If these changes are substantial, it may be necessary to go back to step one and reformulate the interpretive argument. Otherwise, we can go on to step four.

After the first three steps, which constitute the formative stage of the validity argument, the interpretive argument should be reasonably well defined. Steps 4, 5, and 6 involve empirical tests of the interpretive argument and correspond to the summative stage of the validity argument.

Step 4: Identify potential weaknesses in the argument. The aim of the fourth step is to identify the most ploblematic assumptions in the

interpretive argument. Presumably, obvious weaknesses in the interpretive argument would have been identified and, if possible, corrected during steps 1, 2 and 3. The weaknesses identified in step 4 are likely to involve assumptions that are not easily checked using available data. External criticism may be particularly helpful in identifying "hidden" assumptions in the interpretive argument.

Step 5: Conduct empirical studies to check on the most problematic assumptions identified in Step 4. In most cases, the empirical testing of the interpretive argument will involve the collection of data relevant to specific assumptions. If the data tend to refute some of the assumptions, it may be necessary to go back to step one and revise the data collection procedures or the interpretive argument. If the results of the empirical tests support the assumptions under investigation or suggest only minor revisions in these assumptions, we can go on to Step 6 after making any necessary revisions.

Step 6: Evaluate the new argument resulting from Steps 1 to 5. If all of the assumptions and inferences in the interpretive argument seem unproblematic in the context in which they operate, the validity of the interpretation can be accepted, at least for the present. If the argument is not good enough, we may need to go back to Step 4 or to Step 1. It may take several iterations to develop an acceptable interpretive argument with acceptable validity evidence, and even then, a new challenge to some part of the interpretive argument reopens the question of validity.

This six-step process would tend to strengthen the validity argument by eliminating problematic assumptions or by making these assumptions less problematic. In some cases this may be accomplished simply by finding evidence to support the assumption. In other cases it may be necessary to change the interpretation or the measurement procedures so that the

problematic assumption either is not necessary or is at least less problematic.

The checking of assumptions can go on forever if we choose to let it. The "Cheshire Cat" advised Alice to "begin at the beginning, work your way through to the end, and stop". We are not so fortunate as Alice; there is no definite end to be reached. However, we can decide at some point that we have addressed all of the highly problematic assumptions and that the remaining assumptions in the interpretive argument are not particularly problematic. While any of the assumptions could be challenged at any time and could, therefore, become "problematic," we can get to the point where we decide that the argument is good enough for the present and focus our attention on other issues.

On the Advantages of an Argument-based Approach to Validation

This paper opened with the suggestion that while a high degree of consensus has been reached on many issues related to validity, specific guidance on how to evaluate the validity of an interpretation is less readily available. The argument-based approach to validation provides a basis for deciding on the kinds of evidence needed to validate a particular interpretation. It is an attempt to move toward a technology of validation.

An argument-based approach offers several advantages. First, it can be applied to any type of test interpretation or use--the argument-based approach is highly tolerant. It does not discourage the development of any kind of interpretation. It does not preclude the use of any kind of data collection technique in developing a measurement procedure. It does not identify any kind of validity evidence as being generally preferable to any other kind of validity evidence. It does suggest that the interpretation be stated as

clearly as possible, that the interpretation and the test should be consistent with each other, and that the validity evidence should "fit" (be consistent with) the interpretation.

Second, the argument-based approach to validation provides definite guidelines for systematically evaluating the validity of proposed interpretations and uses of test scores. One begins by developing an interpretive argument (a conjecture) and, at each stage of research, one examines those parts of the argument that seem most problematic given previous research and current criticism. Having a clear place to begin and a direction to follow may help to focus serious attention on validation.

Third, although the validation program does not lead to any absolute decision about validity, it does provide a way to gauge progress. As the most problematic inferences and their supporting assumptions are checked and are either supported by the evidence or are adjusted so that they are unproblematic, or at least less problematic, the reasonableness of the interpretive argument as a whole can improve. In some cases, this process may uncover serious flaws that cannot be corrected, and it may, therefore, make sense to abandon the enterprise. In most cases, however, we can expect (or at least hope) that the validity of the interpretation will gradually improve as we eliminate weaknesses in the interpretive argument, and that this improvement will be evident in the clarity and cogency of the argument.

Fourth, the approach may increase the chances that research on validity will lead to improvements in measurement procedures. To the extent that the argument-based approach focuses attention on specific parts of the interpretive argument and on specific aspects of measurement procedures, evidence indicating the existence of a problem (e.g., inadequate coverage of content, the presence of some form of systematic error) may also suggest ways

to solve the problem, and thereby to improve the procedure.

Fifth, the approach is unified in the positive sense that it always involves the development of a preliminary positive case for the interpretive argument and the testing of the assumptions and inferences in the interpretive argument against likely alternatives. It is also unified in the negative sense that it is inconsistent with the view that there are different types of validity that can be used to satisfy the validity requirement; rather, the specific mix of validity evidence needed in each case depends on the interpretations proposed, on the procedures used to collect data, and on the context.

The approach developed here is similar to what Cronbach calls the strong program of construct validation: "a construction made explicit, a hypothesis deduced from it, and pointedly relevant evidence brought in" (Cronbach, 1989, p. 162). The term "argument-based approach to validity" has been used here instead of "construct validity" or the "strong program of construct validity" to emphasize the generality of the argument-based approach, applying as it does to theoretical constructs as well as to attributes defined in terms of specific content or performance domains. Construct validity has often been associated with theory-based interpretations (Cronbach and Meehl, 1955) and therefore the use of this term may be interpreted as suggesting that interpretations that are not closely tied to a theory are inferior to those identified with a specific theory. Interpretive arguments may be, but do not have to be, based on theories. Interpretive arguments can take many different forms; the only restriction is that the claims made about possible interpretations and uses be stated clearly enough to be evaluated.

The expression "argument-based approach" offers some advantages. It is an "approach" to validity rather than a type of validity. By emphasizing the

importance of specifying the interpretive arguments, this terminology highlights the importance of evaluating assumptions, implicit and explicit. The term, "argument", emphasizes the existence of an audience to be persuaded, the need to develop a positive case for the proposed interpretation, and the need to consider and evaluate counterhypotheses.

The argument-based approach to validation does not ensure that more "good works" will be done in the name of validity. However, by identifying the work that needs to be done and by providing a basis for recognizing progress, it could encourage "good works".

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), Test validity (pp. 19-32). Hillsdale, NJ: Lawrence Erlbaum.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing.

Cronbach L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1982). Designing evaluations of educational and social programs. San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.) Test validity (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), Intelligence: Measurement, theory, and public policy. Urbana, IL: University of Illinois Press.

Cronbach, L. J., & Gleser, G. C. (1965). Psychological tests and personnel decisions. Urbana, IL: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Cronbach, L. J., & Snow, R. E. (1977). Aptitudes and instructional methods. New York: Irvington Publishers.

Ebel, R. (1961). Must all tests be valid? American Psychologist, 16, 640-647.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.). Educational Measurement (3rd ed.). New York: American Council on Education and Macmillan.

Frisbie, D. A. (1982). Methods of evaluating course placement systems. Educational Evaluation and Policy Analyses, 4, 133-140.

Guion, R. M. (1974). Open a window: Validities and values in psychological measurement. American Psychologist, 29, 287-296.

Guion, R. M. (1977). Content validity--The source of my discontent. Applied Psychological Measurement, 1, 1-10.

Guion, R. M. (1980). On trinitarian conceptions of validity. Professional Psychology, 11, 385-398.

House, E. R. (1977). The logic of evaluation argument. Los Angeles: Center for the Study of Evaluation.

House, E. R. (1980). Evaluating with validity. Beverly Hills, CA: Sage Publications.

Kane, M. T. (1982). A sampling model for validity. Applied Psychological Measurement, 6, 125-160.

Lakatos, I. (1978). The methodology of scientific research programs. Cambridge, England: Cambridge University Press.

Lakatos, I., & Musgrave, A. (1970). Criticism and the growth of knowledge. Cambridge, England: Cambridge University Press.

Meehl, P. E. (1950). On the circulatory of the law of effect. Psychological Bulletin, 47, 52-75.

Meehl, P. E., & Golden, R. E. (1982). Taxometric methods. In P. Kendall & J. Butcher (Eds.), Handbook of research methods in clinical psychology. New York: Wiley.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. Educational Researcher, 10, 9-20.

Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Wainer and H. Braun (Eds.), Test validity (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed.). New York: American Council on Education and Macmillan.

Nagel, E. (1971). Theory and observation. In E. Nagel, S. Bromberger, & A. Gumbaum, Observation and theory in science. Baltimore: The Johns Hopkins Press

Nedelsky, L. (1965). Science teaching and testing. New York: Harcourt, Brace and World.

Popper, K. R. (1965). Conjecture and refutation: The growth of scientific knowledge. New York: Harper & Row.

Popper, K. R. (1968). The logic of scientific discovery. New York: Harper & Row.

Sawyer, R. (1989). Validating the use of ACT Assessment scores and high school grades for remedial course placement in college (ACT Research Report Series, 89-4). Iowa City, IA: American College Testing.

Snow, R. E. & Lohman, D. E. (1984). Toward a theory of cognitive aptitude for learning from instruction. Journal of Educational Psychology, 76, 347-376.

Snow, R. E. & Lohman, D. E. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational Measurement (3rd ed.). New York: American Council on Education and Macmillan.

Tenopyr, M. L. (1977). Content-construct confusion. Personnel Psychology, 30, 47-54.

Tryon, R. C. (1957). Reliability and behavior domain validity: reformulation and historical critique. Psychological Bulletin, 54, 229-249.

Willingham, W. (1974). College placement and exemption. New York: College Entrance and Examination Board.

Willingham, W. (1988). Testing handicapped people - The validity issue. In H. Wainer & H. Braun (Eds.), Test validity. Hillsdale, NJ: Lawrence Erlbaum.

Table 1

Glossary

Interpretation (of a test score): The meaning/significance assigned to test scores. The interpretation includes statements and/or decisions about the objects of measurement based on the test scores.

Interpretive Argument: The reasoning, implicit or explicit, involved in assigning an interpretation to test scores. The interpretive argument consists of inferences and assumptions leading from test scores to the statements and decisions included in the interpretation.

Problematic Assumptions: Assumptions that are questionable in the context in which the interpretation is being proposed.

Unproblematic Assumptions: Assumptions that are taken as given in the context in which the interpretation is being proposed.

Validity (of an interpretation): The extent to which the interpretive argument supporting the interpretation is plausible and appropriate.

Validity Argument (for an interpretation): The rationale for accepting the inferences and the assumptions in the interpretive argument, based on empirical data, "common sense" and previous research, and quantitative and/or qualitative reasoning. The validity argument provides a basis for accepting the validity of the interpretation.